A Computational Analysis of Complex Noun Phrases in Navy Messages

Elaine Marsh

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory - Code 7510
Washington, D.C. 20375

ABSTRACT

Methods of text compression in Navy messages are not limited to sentence fragments and the omissions of function words such as the copula *be*. Text compression is also exhibited within "grammatical" sentences and is identified within noun phrases in Navy messages. Mechanisms of text compression include increased frequency of complex noun sequences and also increased usage of nominalizations. Semantic relationships among elements of a complex noun sequence can be used to derive a correct bracketing of syntactic constructions.

I INTRODUCTION

At the Navy Center for Applied Research in Artificial Intelligence, we have begun computer-analyzing and processing the compact text in Navy equipment failure messages, specifically equipment failure messages about electronics and data communications systems. These messages are required to be sent within 24 hours of the equipment casualty. Narrative remarks are restricted to a length of no more than 99 lines, and each line is restricted to a length of no more than 69 characters. Because hundreds of these messages are sent daily to update ship readiness data bases, automatic procedures are being implemented to handle them efficiently. Our task has been to process them for purposes of dissemination and summarization, and we have developed a prototype system for this purpose. To capture the information in the narrative, we have chosen to use natural language understanding techniques developed at the Linguistic String Project [Sager 1981].

These messages, like medical reports [Marsh 1982] and technical manuals [Lehrberger 1982], exhibit properties of text compression, in part due to imposed time and length constraints. Some methods of compression result in sentences that are usually called *ill-formed* in normal English texts [Eastman 1981]. Although unusual in normal, full English texts, these are characteristic of messages. Recent work on these properties include discussions of omissions of function words such as the copula *be*, which results in sentence fragments and omissions of articles in compact text [Marsh 1982, 1983; Bachenko 1983]. However, compact text also utilizes mechanisms of compression that are present in normal English but are used with greater frequency in messages and technical reports. Although the messages contain sentence fragments, they also contain many complete sentences. These sentences are long and complicated in spite of the telegraphic style often used. The internal structure of noun phrases in these constructions is often quite complex, and it is in these noun phrases that we find syntactic constructions characteristic of text compression. Similar properties have been noted in other report sublanguages [Lehrberger, 1982; Levi, 1978].

When processing these messages it becomes important to recognize signs of text compression since the function words that so often direct a parsing procedure and reduce the choice of possible constructions are frequently absent. Without these overt markers of phrase boundaries, straightforward parsing becomes difficult and structural ambiguity becomes a serious problem. For example, sentences (1)-(2) are superficially identical, however in Navy messages, the first is a request for a part (an *antenna*) and the second a sentence fragment specifying an antenna performing a specific function. (a *transmit antenna*).

(1) Request antenna shipped by fastest available means.

(2) Transmit antenna shipped by fastest available means.

The question arises of how to recognize and capture these distinctions. We have chosen to take a sublanguage, or domain specific, approach to achieving correct parses by specifying the types of possible combinations among elements of a construction in both structural and semantic terms.

This paper discusses a method for recognizing instances of textual compression and identifies two types of textual compression that arise in standard and sublanguage texts: complex noun sequences and nominalizations. These are both typically found in noun phrase constructions. We propose a set of semantic relations for complex noun sequences, within a sublanguage analysis, that permits the proper bracketing of modifier and host for correct interpretation of noun phrases.

II TEXT COMPRESSION IN NOUN PHRASES

We can recognize the sources of text compression by two means: (1) comparing a full grammar of the standard language to that of the domain in which we are working,

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 1984		**00-00-1984 to 00-00-1984**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Computational Analysis of Complex Noun Phrases in Navy Messages	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory,Navy Center for Applied Research in Artificial Intelligence,Code 7510,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the 22nd Annual Meeting of the Association for Computational Linguistics, held 2-6 July, 1984 in Stanford, CA

14. ABSTRACT
Methods of text compression in Navy messages are not limited to sentence fragments and the omissions of function words such as the copula be. Text compression is also exhibited within ~grammatieal" sentences and is identified within noun phrases in Navy messages. Mechanisms of text compression include increased frequency of complex noun sequences and also increased usage of nominalizations. Semantic relationships among elements of a complex noun sequence can be used to derive a correct bracketing of syntactic constructions.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

and (2) comparing the distribution of constructions in two different sublanguages. The first comparison distinguishes those constructions that are peculiar to a sublanguage [cf. Marsh 1982]. A comparison of a full grammar with two sublanguage grammars, the equipment failure messages discussed here and a set of patient medical histories, disclosed that the sublanguage grammars were substantially smaller than full English grammars, having fewer productions and reflecting a more limited range of modifiers and complements [Grishman 1984]. The second comparison identifies the types of constructions that exhibit text compression. These are common even in full sentences. For example, we found that similar sets of modifiers were used in the two different sublanguages [Grishman 1984]. However, the equipment failure messages had significantly more left and right modifier constructions than the medical, even though the equipment failure messages had about one-half the number of sentences of the patient histories. 236 sentences in the medical domain were analyzed and 123 in the Navy domain. The statistics are presented in Tables 1 and 2.

In particular, there were significantly more noun modifiers of nouns constructions (Noun + Noun constructions) in the equipment failure messages than there were in the medical records, and more prepositional phrase modifiers of noun phrases. Further analysis suggested these constructions are symptomatic of two major mechanisms text compression in Navy messages: of *complex noun sequences* and *nominalizations*.

Complex noun sequences. A major feature of noun phrases in this set of messages is the presence of many long sequences of left modifiers of nouns, (3).

(3) (a) forward kingpost sliding padeye unit
(b) coupler controller standby light
(c) base plate insulator welds
(d) recorder-reproducer tape transport
(e) nbsv or ship-shore tty sat communications
(f) fuze setter extend/retract cycle

Complex noun sequences like these can cause major problems in processing, since the proper bracketing requires an understanding of the semantic/syntactic relations between the components. [Lehrberger 1982] identifies similar sequences (*empilage*) in technical manuals. As he notes, this results from having to give highly descriptive names to parts in terms of their function and relation to other parts.

Modifiers of nouns include nouns and adjectives. In

Left Modifiers of Nouns		
Type	Navy	Medical
Total noun phrases	339	532
Articles	27	38
Adjectival Modifiers:		
Adj	72	136
Adj + Adj	4	34
Possessive N	4	0
Noun Modifiers:		
Noun	99	76
N + N	25	4
Verb	7	0

Table 1: Left Modifier Statistics

Right Modifiers of Nouns		
Type	Navy	Medical
Prepositional Phrases	95	107
Relative Clauses	1	5
Adverb	4	0
Reduced Relative Clauses	7	9

Table 2: Right Modifier Statistics

the sublanguage of Navy messages, unmarked verb modifiers of nouns also occur. This construction is not common in standard English or in the medical record sublanguage mentioned above. It is illustrated above in (2) and below in (4).

(4) (a) receive sensitivity
(b) operate mode
(c) transmit antenna

Because the verbs are unmarked for tense or aspect, they can be mistaken by the parsing procedure for imperative or present tense verbs. Furthermore, in this domain the problem is compounded by the frequent use of sentence fragments consisting of a verb and its object, with no subject present (1) repeated as (5) below.

(5) Request antenna...

Complex noun sequences also commonly arise from the omission of prepositions from prepositional phrases. The resulting long sequences of nouns are not easily bracketed correctly. In this data set, the omission of prepositions is restricted to place and time sequences (6-7).

(6) Request NAVSTA Guantanamo Bay Cuba coordinate ...
Request RSG Mayport arrange....

(7) Original antenna replaced by outside contractor through RSG Mayport 7 JUN 82.

In (6), prepositions marking time phrases have been omitted, and in (7) both time and place prepositions have been omitted.

Nominalizations. The increased frequency of prepositional modifiers in the equipment failure messages was traced to the frequent use of nominalizations in Navy messages. Out of a preliminary set of 89 prepositional modifiers of nouns, 42 were identified as arguments to nominalized verbs (47%), the other 52% were attributive. Examples of argument prepositional phrases are given in (8), attributive in (9).

(8) (a) assistance from MOTU 12
(b) failure of amplifier
(c) cause of casualty
(d) completion of assistance

(9) (a) short circuit between amplifier and power supply
(b) short in cable
(c) receipt NLT 4 OCT 82
(d) burned spots on connector

In these texts, in which nominalization serves as an important mechanism of text compression, it therefore becomes important to distinguish prepositional phrases that serve as arguments of nominalizations from attributive ones.

The syntax of complex modifier sequences in noun phrases and the identification of nominalizations, both characteristic of text compression, need to be consistently defined for a proper understanding of the text being processed. By utilizing the semantic patterns that are derived from a sublanguage analysis, it becomes possible to properly bracket complex noun phrases. This is the subject of the next section.

III SEMANTIC PATTERNS IN COMPLEX NOUN SEQUENCES

Noun phrases in the equipment failure messages typically include numerous adjectival and noun modifiers on the head, and additional modifier types that are not so common in general English. The relationships expressed by this stacking are correspondingly complex. The sequences are highly descriptive, naming parts in terms of their function and relation to other parts, and also describing the status of parts and other objects in the sublanguage. Domain specific information can be used to derive the proper bracketing, but it is first necessary to identify the modifier-host semantic patterns through a distributional analysis of the texts. The basis for sublanguage work is that the semantic patterns are a restricted, limited set. They talk about a limited number of classes and objects and express a limited number of relationships among these objects. These objects and relationships are derived through distributional analysis. and can ultimately be used to direct the parsing procedure.

Complex noun sequences. Semantic patterns in complex noun phrases fall into two types: part names and other noun phrases. Names for pieces of equipment often contain complex noun sequences, i.e. stacked nouns. The relationships among the modifiers in the part names may indicate one of several semantic relations. They may indicate the levels of components. For example, assembly/component relationships are expressed. In *circuit diode*, *diode* is a component of a *circuit*. In *antenna coupler*, *coupler* is a component part of an *antenna*. Part names may also describe the function of the piece of equipment. For example, in the phrase *high frequency transmit antenna*, *transmit* is the function of the *antenna*. The semantic relations among the modifiers of a part are strictly ordered are shown in (10a); examples are provided in (10b).

(10) (a) ID REPAIR SIGNAL FUNCTION PART.

(b) *CU-2007 antenna coupler; HF XMIT antenna; deflection amplifier; UYA-4 display system; primary HF receive antenna*

The component relations in part names are especially closely bound and are best regarded as a unit for processing. Thus *antenna coupler* in *CU-2007 antenna coupler* can be considered a unit. We would not expect to find *antenna CU-2007 coupler* or *coupler CU-2007 antenna*.

In other noun phrases, i.e. those that are not part names, the head nouns can have other semantic categories. For example, looking back at the sentences in (3), the head noun of a noun sequence can be an equipment part (*unit, light*), a process that is performed on electrical signals (*cycle*), a part function (*communica-*

tions). In addition, it can be a repair action (*alignment, repair*), an assistance actions (*assistance*), and so on. Only modifiers with appropriate semantic and syntactic category can be adjoined. For example, in the phrase *fuze setter extend/retract cycle*, semantic information is necessary to attain the correct bracketing. Since only function verbs can serve as noun modifiers, *extend/retract* can be analyzed as a modifier of *cycle*, a process word. *Fuze setter*, a part name, can be treated as a unit because noun sequences consisting of part names are generally local in nature. *Fuze setter* is prohibited from modifying *extend/retract*, since verb modifiers do not themselves take noun modifiers.

Other problems, such as the omissions of prepositions resulting in long noun sequences (cf. (8) and (9) above), can also be treated in this manner. By identifying the semantic classes of the noun in the object of the prepositionless prepositional phrase and its host's class, the occurrence of these prepositionless phrases can be restricted. The date and place strings can then be properly treated as a modifier constructions instead as head nouns.

IV CONCLUSION

Methods of text compression are not limited to omissions of lexical items. They also include mechanisms for maximizing the amount of information that can be expressed within a limited time and space. These mechanisms include increased frequency of complex noun sequences and also increased usage of nominalizations. We would expect to find similar methods of text compression in other types of scientific material and message traffic. The semantic relationships among the elements of a noun phrase permit the proper bracketing of complex noun sequences. These relationships are largely domain specific, although some patterns may be generalizable across domains [Marsh 1984].

The approach taken here for Navy messages, which uses sublanguage selectional patterns for disambiguation, was developed, designed, and implemented initially at the New York University Linguistic String Project for medical record processing [Friedman 1984; Grishman 1983; Hirschman 1982]. It was implemented with the capability for transfer to other domains. We anticipate using a similar mechanism, based partially on the analysis presented here, on Navy messages in the near future.

Acknowledgments

This research was supported by the Office of Naval Research and the Office of Naval Technology PE-62721N. The author gratefully acknowledges the efforts of Joan Bachenko, Judy Froscher, and Ralph Grishman in processing the initial corpus of Navy messages, and the efforts of the researchers at New York University in processing the medical record corpus.

References

[Bachenko 1983] Bachenko, J. and C.L. Heitmeyer. Noun Phrase Compression in Navy Messages. NRL Report 8748.

[Eastman 1981]. Eastman, C.M. and D.S. McLean. On the Need for Parsing Ill-Formed Input. *AJCL* 7 (1981),4.

[Friedman 1984] Friedman, C. Sublanguage Text Processing - Application to Medical Narrative. In [Kittredge 1984].

[Grishman 1983] Grishman, R., Hirschman, L. and C. Friedman. Isolating Domain Dependencies in Natural Language Interfaces. *Proc. of the Conf. on Applied Nat. Lang. Processing (ACL).*

[Grishman 1984] Grishman, R., Nhan, N, Marsh, E. and L. Hirschman. Automated Determination of Sublanguage Syntactic Usage. *Proc. COLING 84*) (current volume).

[Hirschman 1982] Hirschman, L. Constraints on Noun Phrase Conjunction: A Domain-independent Mechanism.*Proc. COLING 82 - Abstracts.*

[Kittredge 1984] Kittredge, R. and R. Grishman.*Proc. of the Workshop on Sublanguage Description and Processing* (held January 19-20, 1984, New York University, New York, New York), to appear.

[Lehrberger 1982]. Lehrberger, J. Automatic Translation and the Concept of Sublanguage. In Kittredge and Lehrberger (eds), *Sublanguage: Studies of Language in Restricted Semantic Domains.* de Gruyter New York, 1982.

[Levi 1978] Levi, J.N. *The Syntax and Semantics of Complex Nominals,* Academic Press, New York.

[Marsh 1982]. Marsh, E. and N. Sager. Analysis and Processing of Compact Text. Proc. COLING 82, 201-206, North Holland.

[Marsh 1983] Marsh, E. Utilizing Domain-Specific Information for Processing Compact Text. Proc. Conf. Applied Natural Language Processing, 99-103 (ACL).

[Marsh 1984] Marsh E. General Semantic Patterns in Different Sublanguages. In [Kittredge 1984].

[Sager 1981] Sager, N. *Natural Language Information Processing.* Addison-Wesley, Reading, MA.